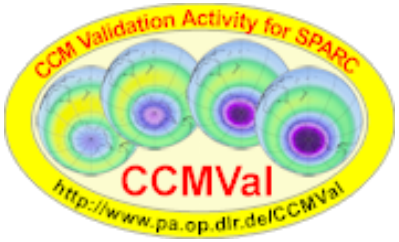
Science & Technology Facilities Council

CCMVal Archive at BADC

Charlotte Pascoe
Bryan Lawrence, Alison Pamment, David Hooper, Sam Pepler

British Atmospheric Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL



CCMVal Archive at BADC

- BADC
- CF-NetCDF
- Three good reasons to keep Data:
  Reuse: in 10 minutes, in 10 years...
  Repurposing: discovery
  Citation: data journals - OJIMS
- CMIP5 and Metafor: Metadata Capture and YOU!

Science & Technology Facilities Council

CEDA

Centre for Environmental Data Archival

- British Atmospheric Data Centre
- NERC Earth Observation Data Centre
- IPCC Data Distribution Centre

- UK Projects
  - •NDG (NERC Data Grid)
  - •C-SEKT (New CSML and CSML-WFS)
  - •OJMS (Data Publication)
  - •UKCIP (UK-Scale Climate Predictions)
- EU Projects
  - •Metafor (Climate Model Documentation)
  - •EUFAR (Distributed Archive for European Aircraft)
  - •IS-ENES (Distributed Archive for European Climate Model Data)

NERC designated data centres for Atmospheric Science and EO

IPPC official data centre http://ipcc-data.org

... and just for the BADC

- BADC has approximately 150 real datasets (and thousands of virtual datasets).
- BADC has tens of millions of files containing thousands of measured or simulated parameters.
- BADC [illegible] scribe tho[illegible] ong wit[illegible] em.
- Calend[illegible]
  - 1[illegible]
  - downloaded 30 TB data
  - in 13 million files
  - from 134 datasets.

CCMVal Data: 3.4 TB

Less than half of the BADC data consumers are "atmospheric science" users!

CF-NetCDF

Facilitating the exchange of data between modelling groups

Standard Names

Each physical, chemical or biological quantity is given a CF standard name and assigned appropriate units

The purpose of the standard name is to determine whether the parameters from different data sources can be treated as directly comparable.

Coordinate Variables

CF metadata describes multi-dimensional data by attaching a coordinate variable to each dimension

Time

A separate variable is used to describe the time dimension

Standard Name	Canonical Units	AMIP	GRIB
air_potential_temperature	K	theta	13
air_pressure	Pa	plev	1

CF-NetCDF

Files composed of just three basic entry types

Global Attributes

store items of metadata which are relevant to the file as a whole

Dimensions

e.g. **time**, **altitude**, **range**

each one has a corresponding coordinate **variable** which stores the coordinate values along the dimension

Variables

contain the actual data

can have **Attributes** attached to them

The first 10 (of 240) lines from the file **sw010203**
(taken from the NERC MST Radar Facility archives)

```
4.31 155.3  3.92 136.1  5.15 140.2  4.23 137.1  4.75 150.2  4.71 137.9
4.35 146.5  4.52 138.0  4.83 153.7  5.40 145.8  4.63 141.0  4.90 137.3
4.31 143.3  4.58 157.0  4.94 141.7  4.65 143.1  4.63 143.0  4.88 149.5
5.42 148.5  4.92 140.4  4.04 146.7  3.92 151.5  5.02 135.3  5.06 151.6
4.65 152.3  4.31 168.8  3.79 145.3  5.92 152.9  5.02 145.8  4.77 161.6
4.79 144.1  4.60 147.5  5.33 150.1  4.81 141.0  6.02 146.9  4.38 149.0
4.42 142.5  4.58 133.4  4.35 150.5  4.96 149.8  5.56 143.4  5.08 148.5
5.19 141.6  4.40 142.4  4.10 152.6  5.02 134.0  4.94 142.9  5.27 144.4
5.38 141.5  5.88 144.8  6.00 140.1  4.75 158.3  5.08 148.1  5.46 163.5
4.27 150.8  4.69 138.8  5.71 144.0  5.21 138.8  5.00 132.4  5.06 144.4
```

What is known about this file?

sw indicates that the file contains "surface" wind data (i.e. speed and direction) from the location Frongoch

010203 represents the date in YYMMDD format

1st February 2003
(British convention)

2nd January 2003
(North American convention)

3rd February 2001
(Swedish convention)

The first 10 (of 240) lines from the file **sw010203**
(taken from the NERC MST Radar Facility archives)

```
4.31 155.3  3.92 136.1  5.15 140.2  4.23 137.1  4.75 150.2  4.71 137.9
4.35 146.5  4.52 138.0  4.83 153.7  5.40 145.8  4.63 141.0  4.90 137.3
4.31 143.3  4.58 157.0  4.94 141.7  4.65 143.1  4.63 143.0  4.88 149.5
5.42 148.5  4.92 140.4  4.04 146.7  3.92 151.5  5.02 135.3  5.06 151.6
4.65 152.3  4.31 168.8  3.79 145.3  5.92 152.9  5.02 145.8  4.77 161.6
4.79 144.1  4.60 147.5  5.33 150.1  4.81 141.0  6.02 146.9  4.38 149.0
4.42 142.5  4.58 133.4  4.35 150.5  4.96 149.8  5.56 143.4  5.08 148.5
5.19 141.6  4.40 142.4  4.10 152.6  5.02 134.0  4.94 142.9  5.27 144.4
5.38 141.5  5.88 144.8  6.00 140.1  4.75 158.3  5.08 148.1  5.46 163.5
4.27 150.8  4.69 138.8  5.71 144.0  5.21 138.8  5.00 132.4  5.06 144.4
```

What can we guess?

- Values are clearly arranged in pairs
  - 1st value of pair (e.g. 4.31) must represent speed - probably in units of m s^{-1}
  - 2nd value of pair (e.g. 155.3) must represent direction - probably in units of ° from North (but meteorological or vector convention?)
- 240 lines, each with 6 columns, each with a pair of values => 1440 pairs of values
- There are 1440 minutes in a day => 1 minute sampling

The first 10 (of 240) lines from the file **sw010203**
(taken from the NERC MST Radar Facility archives)

```
4.31 155.3  3.92 136.1  5.15 140.2  4.23 137.1  4.75 150.2  4.71 137.9
4.35 146.5  4.52 138.0  4.83 153.7  5.40 145.8  4.63 141.0  4.90 137.3
4.31 143.3  4.58 157.0  4.94 141.7  4.65 143.1  4.63 143.0  4.88 149.5
5.42 148.5  4.92 140.4  4.04 146.7  3.92 151.5  5.02 135.3  5.06 151.6
4.65 152.3  4.31 168.8  3.79 145.3  5.92 152.9  5.02 145.8  4.77 161.6
4.79 144.1  4.60 147.5  5.33 150.1  4.81 141.0  6.02 146.9  4.38 149.0
4.42 142.5  4.58 133.4  4.35 150.5  4.96 149.8  5.56 143.4  5.08 148.5
5.19 141.6  4.40 142.4  4.10 152.6  5.02 134.0  4.94 142.9  5.27 144.4
5.38 141.5  5.88 144.8  6.00 140.1  4.75 158.3  5.08 148.1  5.46 163.5
4.27 150.8  4.69 138.8  5.71 144.0  5.21 138.8  5.00 132.4  5.06 144.4
```

In which order should we read the data?

Column by column and then row by row or *vice versa*?

- Try both ways and plot time series of the speed and direction data
- There should be no sharp discontinuities in speed or direction

Vector (i.e. towards which the wind is blowing) or meteorological direction?

- Compare with synoptic pressure maps or MST radar data

The first 10 (of 240) lines from the file **sw010203**
(taken from the NERC MST Radar Facility archives)

```
4.31 155.3  3.92 136.1  5.15 140.2  4.23 137.1  4.75 150.2  4.71 137.9
4.35 146.5  4.52 138.0  4.83 153.7  5.40 145.8  4.63 141.0  4.90 137.3
4.31 143.3  4.58 157.0  4.94 141.7  4.65 143.1  4.63 143.0  4.88 149.5
5.42 148.5  4.92 140.4  4.04 146.7  3.92 151.5  5.02 135.3  5.06 151.6
4.65 152.3  4.31 168.8  3.79 145.3  5.92 152.9  5.02 145.8  4.77 161.6
4.79 144.1  4.60 147.5  5.33 150.1  4.81 141.0  6.02 146.9  4.38 149.0
4.42 142.5  4.58 133.4  4.35 150.5  4.96 149.8  5.56 143.4  5.08 148.5
5.19 141.6  4.40 142.4  4.10 152.6  5.02 134.0  4.94 142.9  5.27 144.4
5.38 141.5  5.88 144.8  6.00 140.1  4.75 158.3  5.08 148.1  5.46 163.5
4.27 150.8  4.69 138.8  5.71 144.0  5.21 138.8  5.00 132.4  5.06 144.4
```

It is often possible to "decode" ASCII files in this way, it is much more difficult for binary.

No-one will be prepared to make this effort unless they have a strong need for the data.

The data will becomes useless if the file name is changed - the date information is not recorded anywhere else.

Even if the data can be read, they may be of little scientific value unless something is known about: the type of instrument used, where it was located & how it was operated.

The partial contents of file nerc-mstrf-wind-sensors_capel-dewi_20080114_wxt510.nc

```
global attributes:
     :verbose_metadata = "Free text description" ;
     :file_version_number = 1s ;
     :data_year = 2008s ;
     :data_month = 1s ;
     :data_day = 14s ;
dimensions:
     time = 1440 ;
variables:
     float longitude() ;
          longitude:units = "degrees_east" ;
          longitude:axis = "X"
     float latitude() ;
          latitude:units = "degrees_north" ;
          latitude:axis = "Y" ;
     float altitude() ;
          altitude:units = "m" ;
          altitude:axis = "Z" ;
     int time(time) ;
          time:units = "seconds since 2008-01-14 00:00:00 +00:00" ;
          time:axis = "T" ;
     float mean_wind_speed(time) ;
          mean_wind_speed:units = "m s-1" ;
          mean_wind_speed:coordinates = "latitude longitude altitude" ;
          mean_wind_speed:cell_methods = "time: minimum (interval: 3 s)" ;
          mean_wind_speed:missing_value = 99.9f ;
     short mean_wind_direction(time) ;
          mean_wind_direction:units = "degree" ;
          mean_wind_direction:coordinates = "latitude longitude altitude" ;
          mean_wind_direction:cell_methods = "time: minimum (interval: 3 s)";
          mean_wind_direction:missing_value = 999s ;
```

We need metadata

We will never *look* at all our data. We need to process it on ingestion.
We *have* to provide tools so users can select what they want
Tools need metadata!

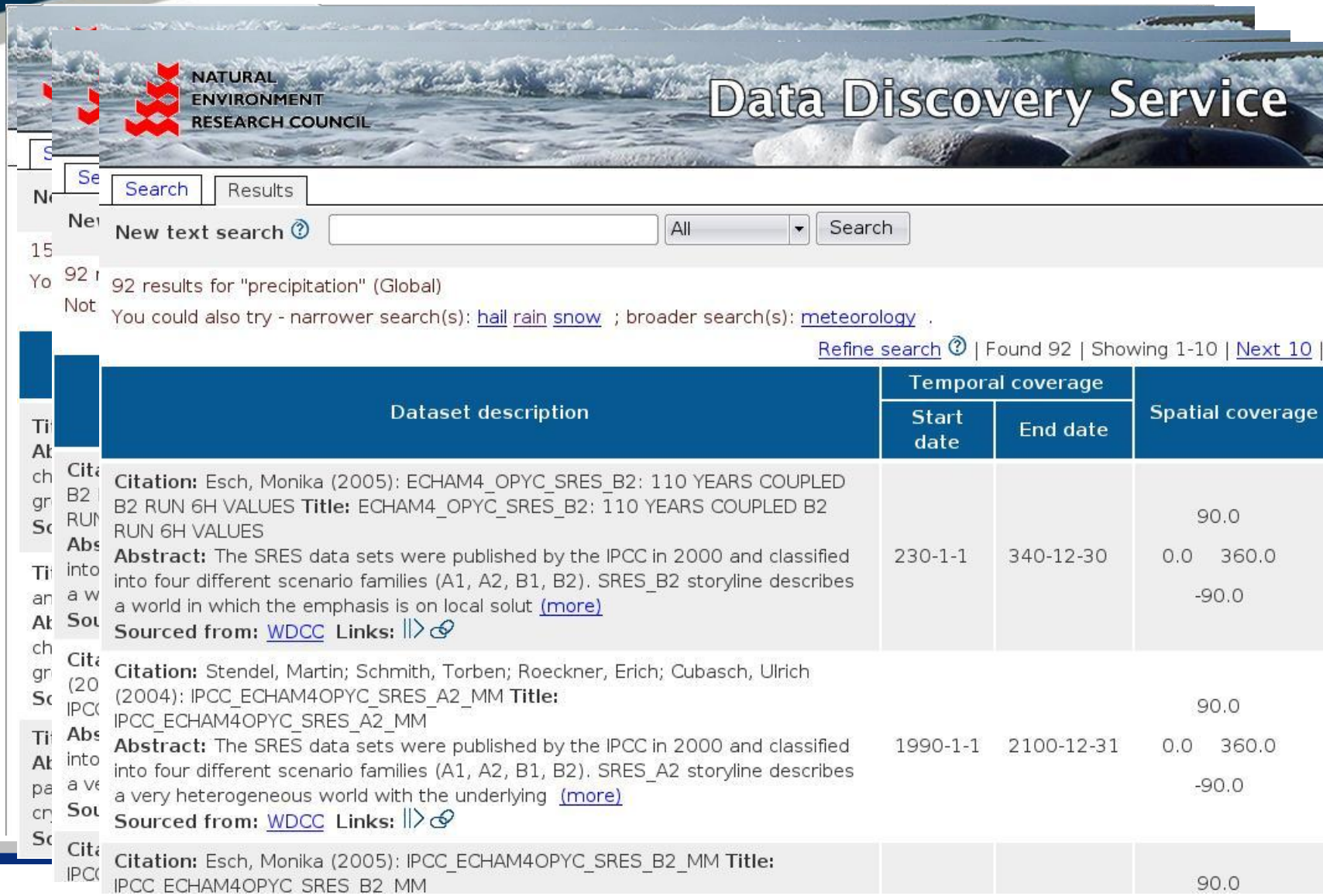
We can't manually reprocess our files to create new information about the data we hold, we *have* to automate
Automation needs compliant metadata

Discovery Choosing Manipulating

All need input metadata to drive tools

Science & Technology Facilities Council

Ontology Mediated Discovery



How discovery works

Vocabulary server for CF standard names

Defines: **Exact** (for previous versions of the standard name), **Narrow** and **Broad** mappings

Rain.....	- Broad Match:	Precipitation
Precipitation	- Narrow Match:	Hail, Rain, Snow
	- Broad Match:	Meteorology

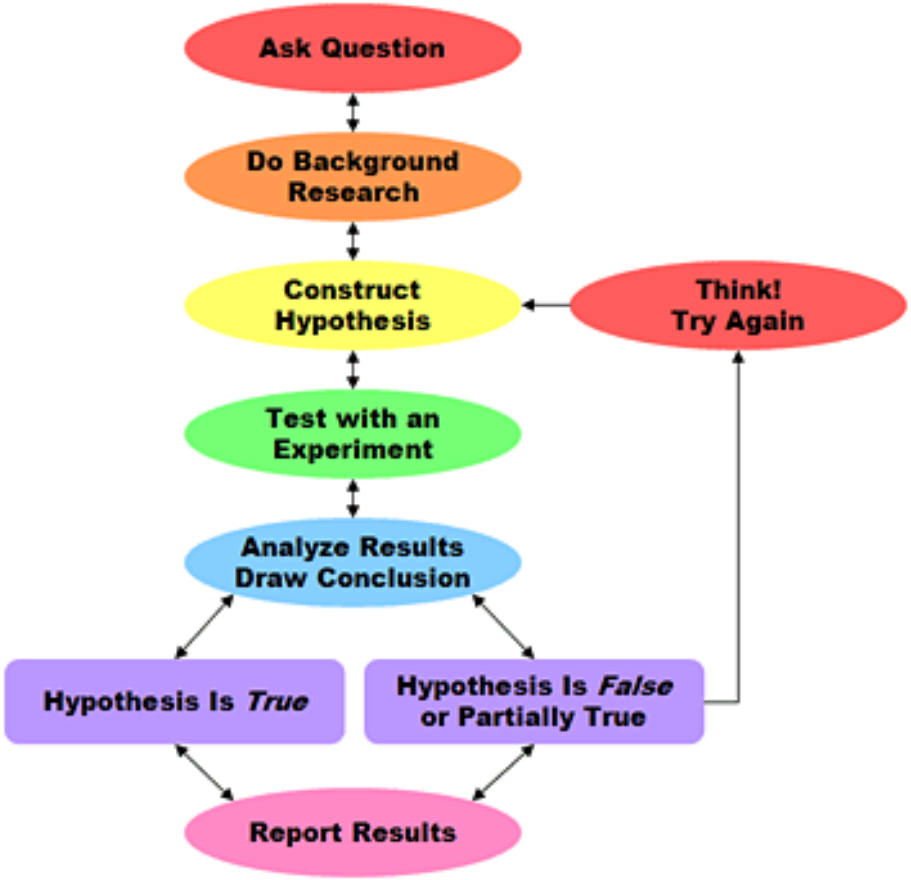
Building the mappings between vocabularies requires understanding the science.

Roy Lowry (BODC) has built the equivalences that are currently in the vocab server but...

More people will need to **get involved** if full mappings are to be provided between all the stored vocabularies.

The Scientific Method

- depends on Analyze Results,
- finishes with REPORT RESULTS!
- Once upon a time, data were a few numbers, maybe a table, and the “paper” included “results”.
- That's very rarely true now, especially in environmental sciences!
- Do we care any more about the evidential basis for the scientific method? Do we care about repeatability?



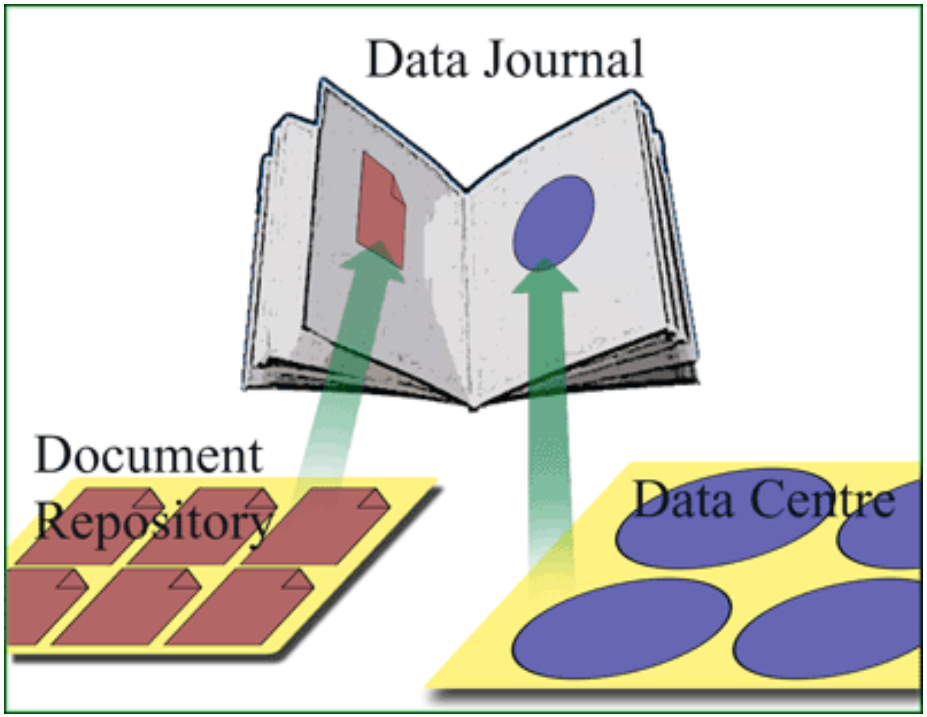
Overlay Journal Infrastructure for Meteorological Sciences

http://proj.badc.rl.ac.uk/ojims

Aid reusability: make data **permanently** available and the knowledge within useable in other contexts

Enable Recognition: make it possible to measure and recognise the value of work

- Author prepares the overlay document.
- Author submits the dataset and the overlay document.
- Reviewer reviews the overlay document and the dataset



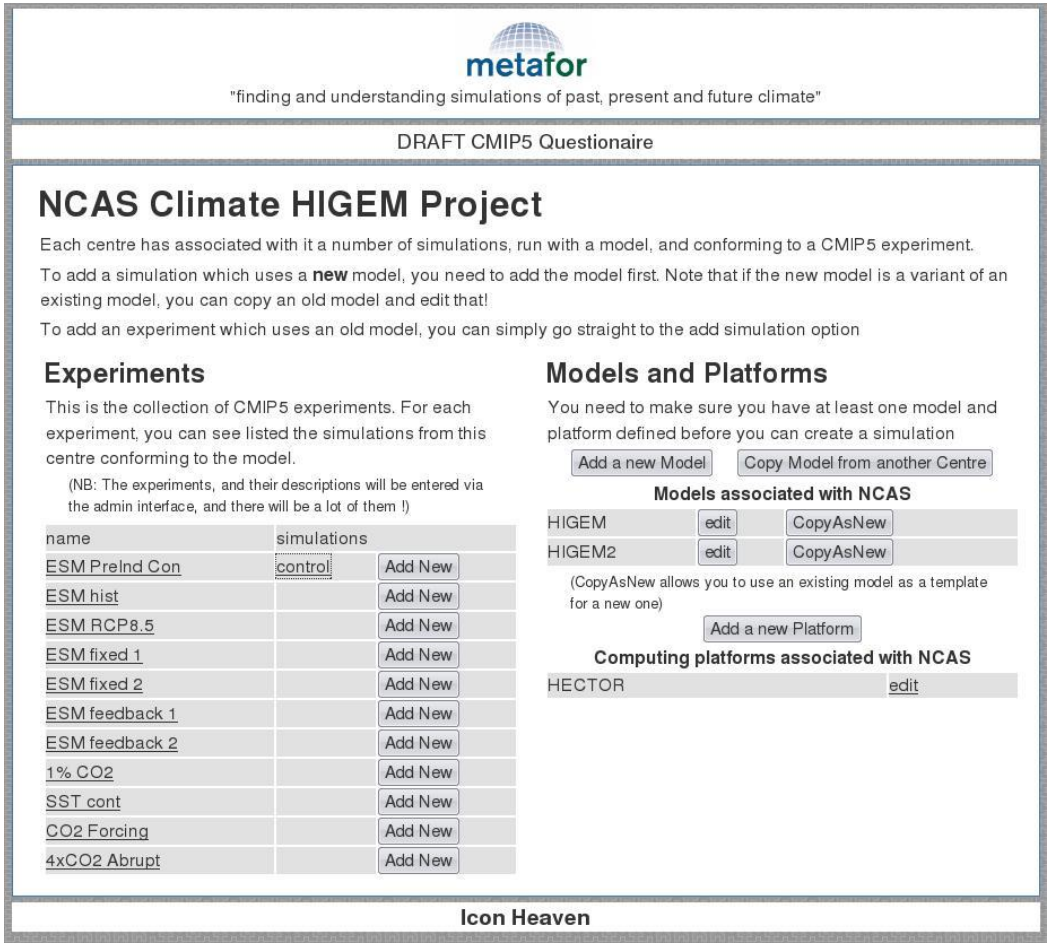
Science & Technology Facilities Council

CMIP5 Metadata Capture

metafor

Collect information from modelling groups:

- Details of models used
- Details of simulations carried out
- Details of how models conform to experiment requirements
- Details of hardware used
- Make information available to Curator, PCMDI, METAFOR and others



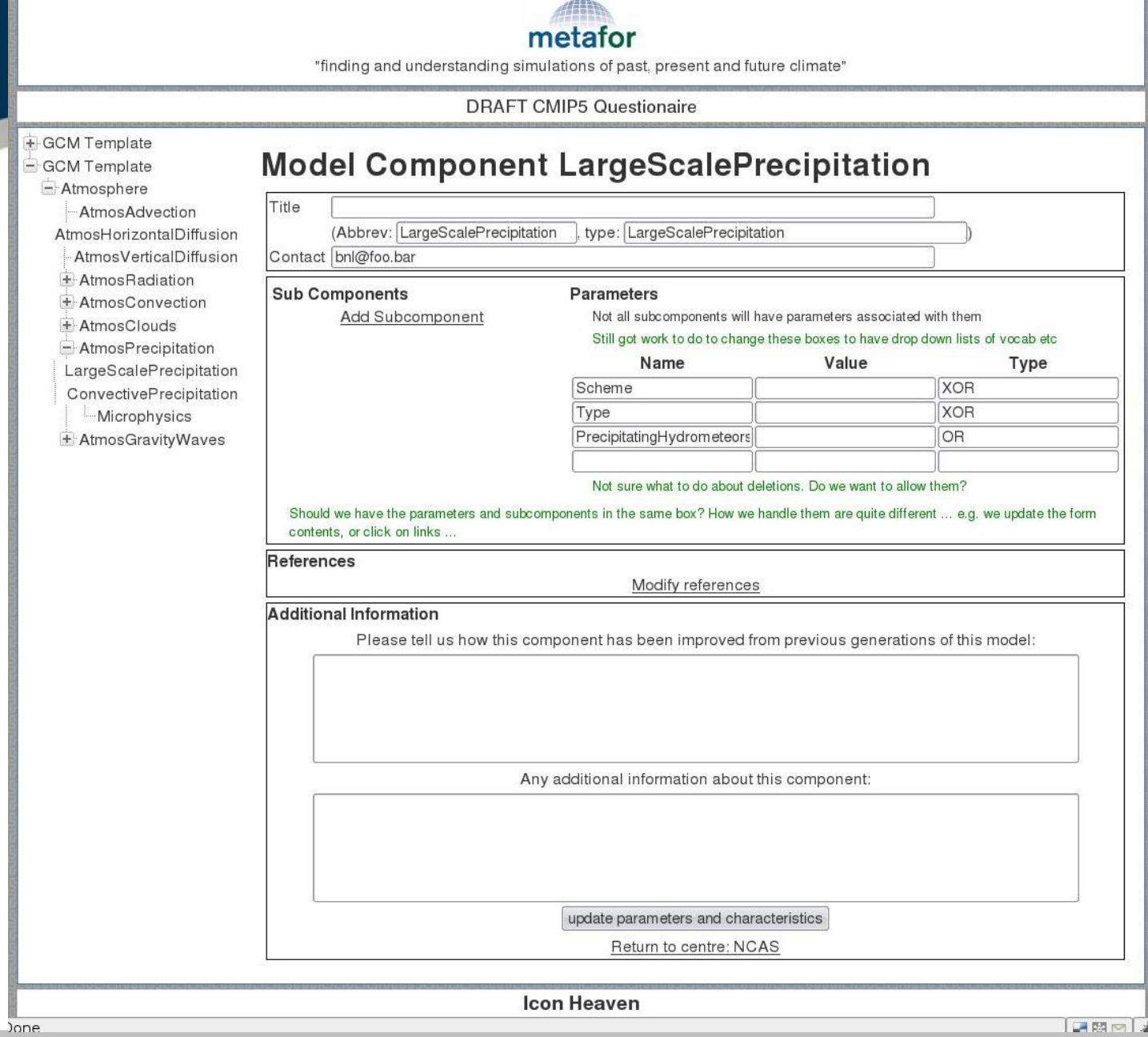
prototype of the questionnaire

Science & Technology Facilities Council

metafor

For CMIP5, the questionnaire will be more flexible than the CCMVal prototype

Users can...
- complete the questionnaire in the order that suits them.
- add their own subcomponents.



British Atmospheric Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL

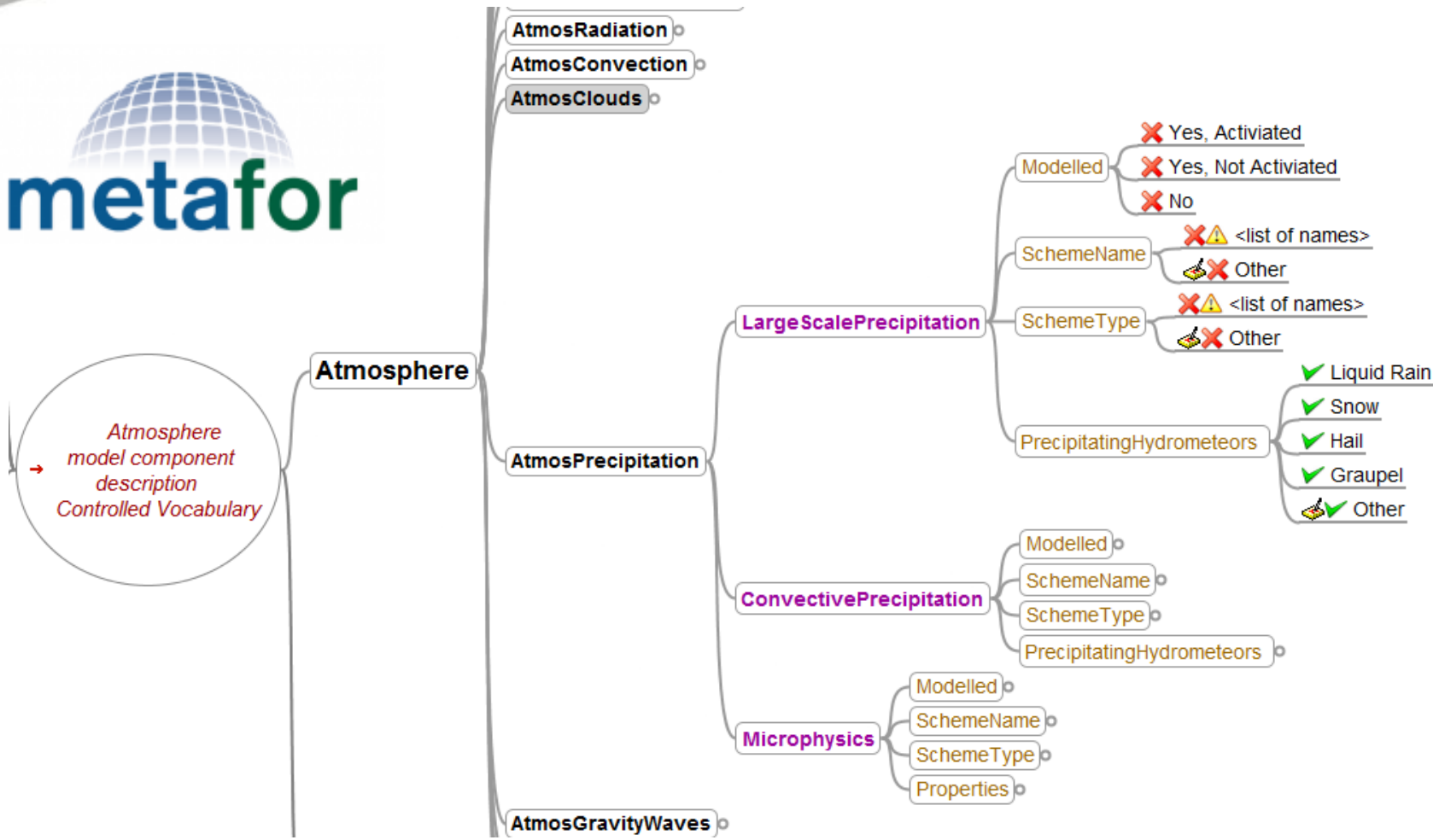
CCMVal

Science & Technology Facilities Council

CMIP5 Metadata Capture

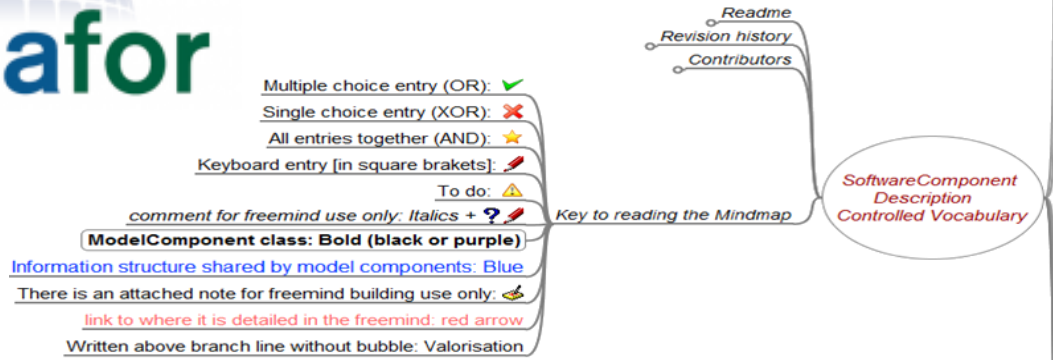
metafor

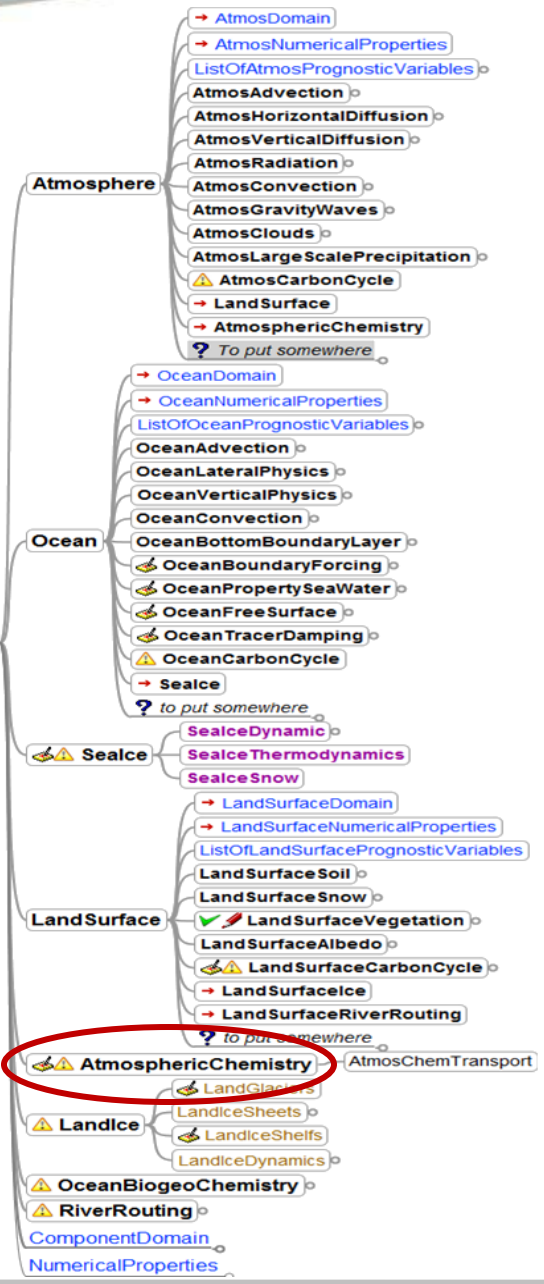
Atmosphere model component description Controlled Vocabulary



Many controlled vocabularies being built: Software

metafor





Working closely with modelling community and Curator partners and..**YOU!**

Thank You

CCMVal Archive at BADC

- BADC
- CF-NetCDF
- Three good reasons to keep Data:
  Reuse: in 10 minutes, in 10 years...
  Repurposing: discovery
  Citation: data journals - OJIMS
- CMIP5 and Metafor: Metadata Capture and YOU!